Position-dependent Codon Usage Bias in the Human Transcriptome

Kaavya Subramanian[1], Nathan Waugh[2], Cole Shanks[1], David A Hendrix[1,2]

1. School of Electrical Engineering and Computer Science, Oregon State University
2. Department of Biochemistry and Biophysics, Oregon State University

ABSTRACT

All life depends on the reliable translation of RNA to protein according to complex interactions between translation machinery and RNA sequence features. While ribosomal occupancy and codon frequencies vary across coding regions, well-established metrics for computing coding potential of RNA do not capture such positional dependence. Here, we investigate *position-dependent codon usage bias (PDCUB)*, which dynamically accounts for the position of protein-coding signals embedded within coding regions. We demonstrate the existence of PDCUB in the human transcriptome, and show that it can be used to predict translation-initiating codons with greater accuracy than other models. We further show that observed PDCUB is not accounted for by other common metrics, including position-dependent GC content, consensus sequences, and the presence of signal peptides in the translation product. More importantly, PDCUB defines a spectrum of translational efficiency supported by ribosomal occupancy and tRNA adaptation index (tAI). High PDCUB scores correspond to a tAI-defined translational ramp and low ribosomal occupancy, while low PDCUB scores exhibit a translational valley and the highest ribosomal occupancy. Finally, we examine the relationship between PDCUB intensity and functional enrichment. We find that transcripts with start codons showing the highest PDCUB are enriched for functions relating to the regulation of synaptic signaling and plasticity, as well as skeletal, heart, and nervous-system development. Furthermore, transcripts with high PDCUB are depleted for functions related to immune response and detection of chemical stimulus. These findings lay important groundwork for advances in our understanding of the regulation of translation, the calculation of coding potential, and the classification of RNA transcripts.

INTRODUCTION

Translation from RNA to protein is a fundamental and ubiquitous life process. Some studies estimate that ≈68% of human genes do not encode proteins, and are transcribed as long noncoding RNAs (lncRNAs) or transcripts of unknown coding potential [1]. This raises the question of how the ribosome distinguishes mRNAs from lncRNAs that have open reading frames (ORFs) [2]. At the same time, other studies suggest that some lncRNAs are weakly translated [3-5]. These observations underscore the importance of understanding how the ribosome uses sequence features to distinguish mRNAs from lncRNAs, as well as to properly identify the start codon within an mRNA. Such sequence features could help explain the translation of short transcripts that encode small peptides [6, 7], and improve the design of mRNA vaccines [8].

The Kozak consensus sequence is one such motif, discovered by Marilyn Kozak in the 1980s [9]. This sequence characterizes and helps identify start codons in eukaryotic mRNAs and is regarded as a key feature that enhances translational potential. Another approach to quantifying mRNA coding potential is codon usage bias (CUB), which refers to the frequency of usage for each codon in the coding portion of a transcriptome relative to the frequency of synonymous codons. Computational models of CUB have been shown to perform well in predicting translational efficiency and also show an association with ribosome profiling data [10]. The codon adaptation index (CAI) is a widely recognized metric of CUB that assigns a score to each transcript based on its length and codon composition relative to overall CUB across the coding transcriptome [11]. The CAI model has been used as a baseline for other quantitative models of CUB, and has also been correlated with gene expression in select genomes [11].

A potential limitation of these approaches is that they do not consider the role of codon position or transcript length explicitly. Recently, studies of position-dependent codon usage bias (PDCUB) observed that codon usage is non-uniform with regard to transcript position in *E. coli* [12] and yeast [13]. Moreover, our own investigations demonstrated that a recurrent neural network, which we trained to distinguish mRNAs from lncRNAs based on sequence alone, was able to independently learn sequence-specific rules and make classification decisions approximately 100-200 nucleotides (nt) downstream of the start codon [14].

Here, we investigate the extent of PDCUB within the human coding transcriptome and whether PDCUB can function as a predictor of start codon location. We present a codon bias score that incorporates these position-dependent codon patterns, using only information observed in the first 300 nt downstream of a given AUG. We show that our score is capable of distinguishing start codons from non-start AUGs with greater accuracy than other methods. While our PDCUB score was constructed from general positional trends in codon usage, we show that it is correlated with translational efficiency data. We show that high PDCUB is associated with a translational ramp and low ribosomal occupancy, and low PDCUB is associated with a translational valley and high ribosomal occupancy. Moreover, we show that high- and low-PDCUB transcripts show a clear division in biological function corresponding to a need for high and low translational efficiency.

RESULTS

Visualization of PDCUB

We first performed a coarse-grained investigation of PDCUB in the human transcriptome using GENCODE data [15]. The frequencies of all 61 sense codons in protein-coding transcripts were visualized in successive 60-nt bins, up to 3000 nt downstream of the annotated start codon. We transformed the observed bin-specific codon frequencies to z-scores, defined relative to the global average frequency of codon occurrences within a bin, across the 50 bins downstream of the AUG. We observed that most codons displayed a distinct position-dependent z-score profile as we move downstream from the start codon, with some codons increasing in usage and others decreasing (Figure 1A). This led us to evaluate a more refined model of PDCUB in the human transcriptome.

We investigated whether the heatmap of z-score profiles reveals other patterns in position-dependent codon usage when the codons are clustered according to distinct criteria. First, we organized codons by their amino acid chemical properties, but did not observe a strong correspondence between z-score profile and encoded amino acid chemical type (Supplementary Figure 1). We next performed hierarchical clustering on the codons and found that they were predominantly organized by G and C (GC) nucleotide content. To test this observation, we compared our computed z-score dendrogram distance against a modified Hamming distance that compares the extent of variation of GC nucleotides between two codons (Figure 1A). We found a strong correspondence between the number of common GC nucleotides and the dendrogram distance in Figure 1A, suggesting that the GC content of each codon is indeed a significant factor in the pattern of position-dependent z-scores (Supplementary Figure 2).

Next, we grouped codons by their GC content and plotted z-score curves for each individual codon (Figure 1B). We found that codons consisting exclusively of GC nucleotides show a strong enrichment near the start codon, while codons consisting exclusively of A or U (AU) nucleotides are depleted immediately after the start codon. Codons with either 1 or 2 GC nucleotides continue this trend, with GC-rich codons (2 GC nucleotides) showing greater similarity to GC-only codons, while GC-poor codons (1 GC nucleotide) show greater similarity to AU-only codons.

Optimizing PDCUB weight matrix

We used the prediction of start codons to assess the importance of the observed PDCUB patterns in the human transcriptome. We created training, validation, and test sets by subdividing the protein-coding transcripts from GENCODE using an 80:10:10 split. We created PDCUB position-specific scoring matrices (PSSMs, or weight matrices) to score the occurrence of a codon *c* at bin *b* of a potential open reading frame, relative to the global frequency of the nucleotide triplet corresponding to the codon *c*. We assigned the PDCUB weight matrix score (PDCUB score) to the series of triplets that follow each AUG in the transcripts across all frames, to discern whether the PDCUB score was sufficient to distinguish start codons from all other AUGs in a transcript.

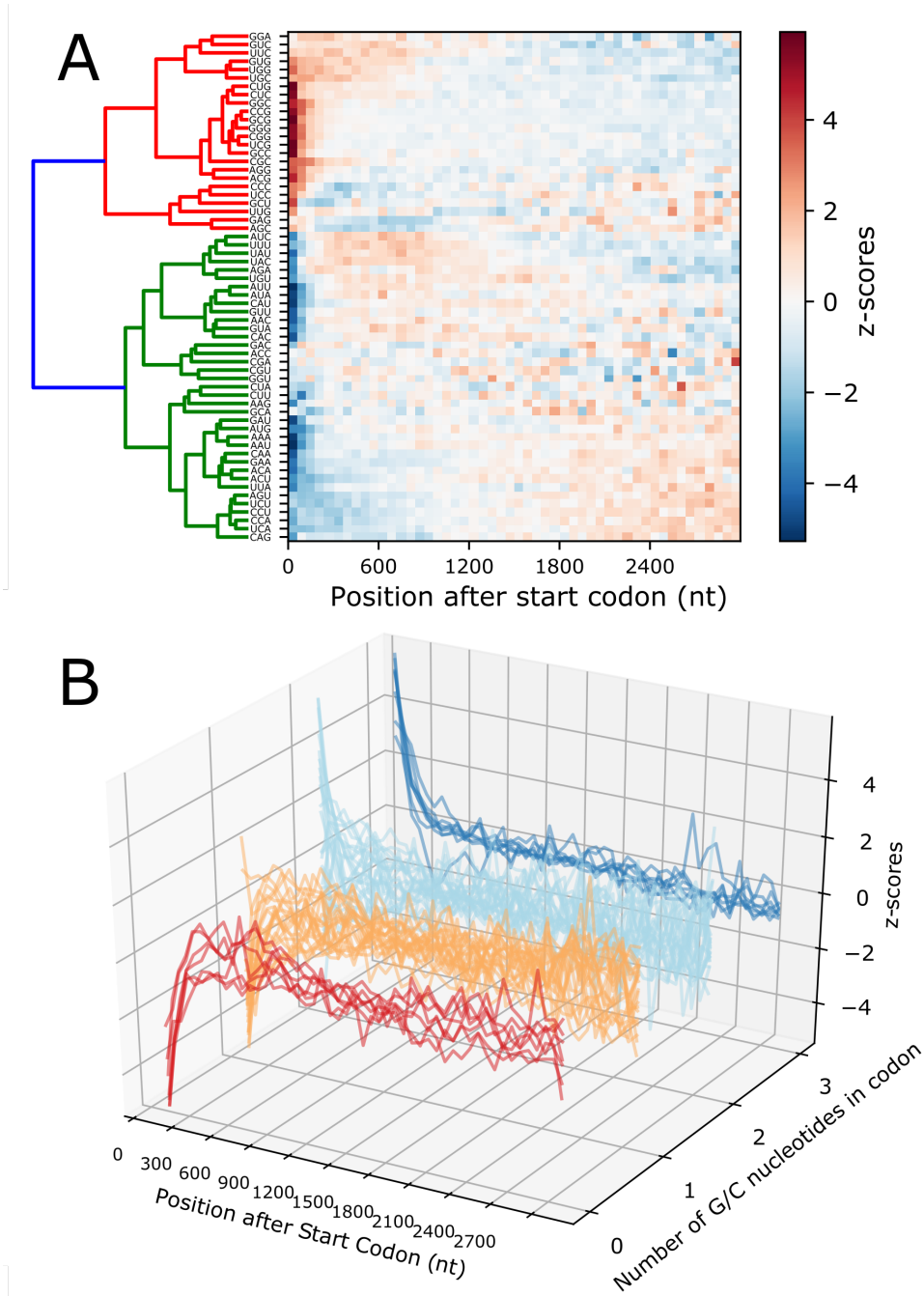

Figure 1: ***A.*** *A "bird's eye view" of position-dependent codon usage bias in the human transcriptome (GENCODE transcript models). A heatmap and dendrogram visualizing the z-scores computed from codon occurrence frequency in bins of size 60nt.* ***B.*** *Curves show the z-scores computed from codon occurrence frequencies, shown in clusters corresponding to the number of G/C nucleotides within the codon.*

We focused on the first 300 nt downstream of the start codon so that our score could be applied to most mRNA and lncRNA transcripts, including those with shorter ORFs. As a metric of predictive performance, we computed a recall, defined as the percentage of transcripts in which our PDCUB score identified the start codon as the highest-scoring AUG in that transcript. This seemed to depend on bin number and bin size, so we systematically computed the recall for combinations of both parameters (Figure 2A). We computed these parameters (bin number and size) for our PDCUB weight matrix using the training set, selected an optimal set of parameters using performance on the validation set, and computed final percentages on the test set. We found that start codons were detected with maximum accuracy using one codon per bin – in other words, the combination of 100 single-codon bins provided the greatest predictive accuracy (Figure 2A). The z-score heatmap for this optimal bin size and number of bins is presented in Figure 2B. Hierarchical clustering of this heat map also corresponds to codon GC-content.

Information content of the PDCUB model

We used the PDCUB probability matrix to compute the information content as a function of nucleotide position or bin (see Methods). We observed enriched information content in the range from codons 0 to 20, relative to and downstream of the start codon (Figure 2C).

Start codon identification as a function of transcript length

We next investigated the role of transcript length in the percentage of correctly identified start codons – specifically, whether transcript length is correlated with the percentage of start codons identified by our PDCUB score as the highest-scoring AUG triplet in a given transcript. We compared the central tendencies (mean and median) for PDCUB scores computed for start codons and for other AUGs. We found that there was a significant difference between start codons and non-start AUGs for all transcript lengths (Supplementary Figure 3).

We compared the length-dependent recall of the PDCUB score with four other predictive models. First, we compared the PDCUB model to a Kozak consensus sequence model, which relies on finding within a transcript the nucleotide sequence of fewest possible mismatches to the Kozak consensus sequence [16]. We also compared the PDCUB model to a weight matrix computed from base composition in the region of the Kozak sequence, from –10 to +1 relative to the start codon. A null model that estimates the rate for the random selection of start codons amongst all AUG triplets, and a position-dependent GC-content model, were also used as controls. We found that the PDCUB model consistently performed best among the five models (Figure 3A). Even though PDCUB only considers the first 300 nt of the coding sequence, we observed improved performance with shorter transcripts. Figure 3B shows an example transcript, plotting the PDCUB score over all positions, and plotted separately for each reading frame. This example transcript conforms to the trend observed in the majority of transcripts of this length, with the start codon having the highest PDCUB score.

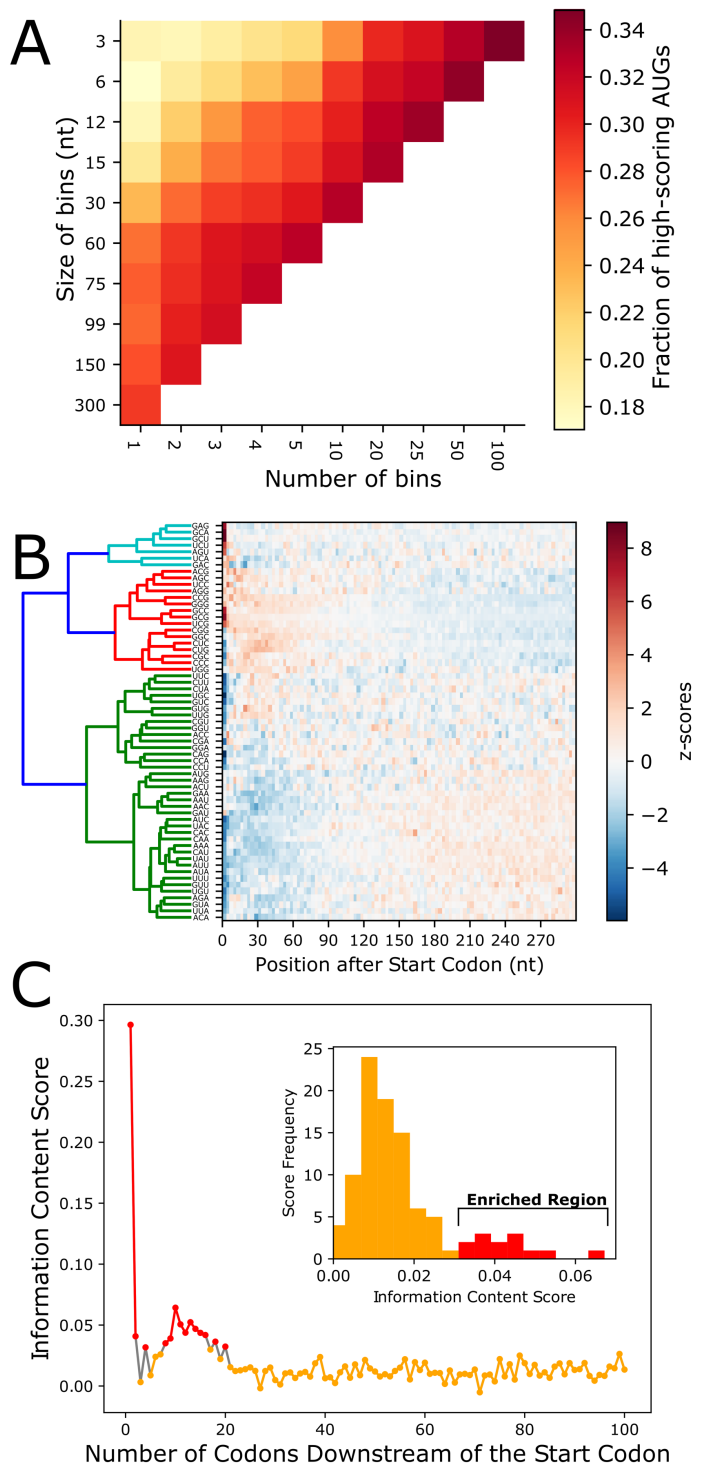

Figure 2: ***A.*** *A heatmap demonstrating the optimal bin size and number of bins within the first 300 nt of CDS regions of the human transcriptome favors 100 3-nt bins.* ***B.*** *A codon z-score heatmap weight hierarchical clustering for the 300 nt after the start codon* ***C.*** *The information content of all transcripts, computed from position-dependent codon frequencies shows a region of enriched information content relative to the background trinucleotide frequencies. Red line segments indicate some regions of enriched information content. The inset shows a histogram of the information content for the first 100 codons with a tail of enriched information content, shown in red.*

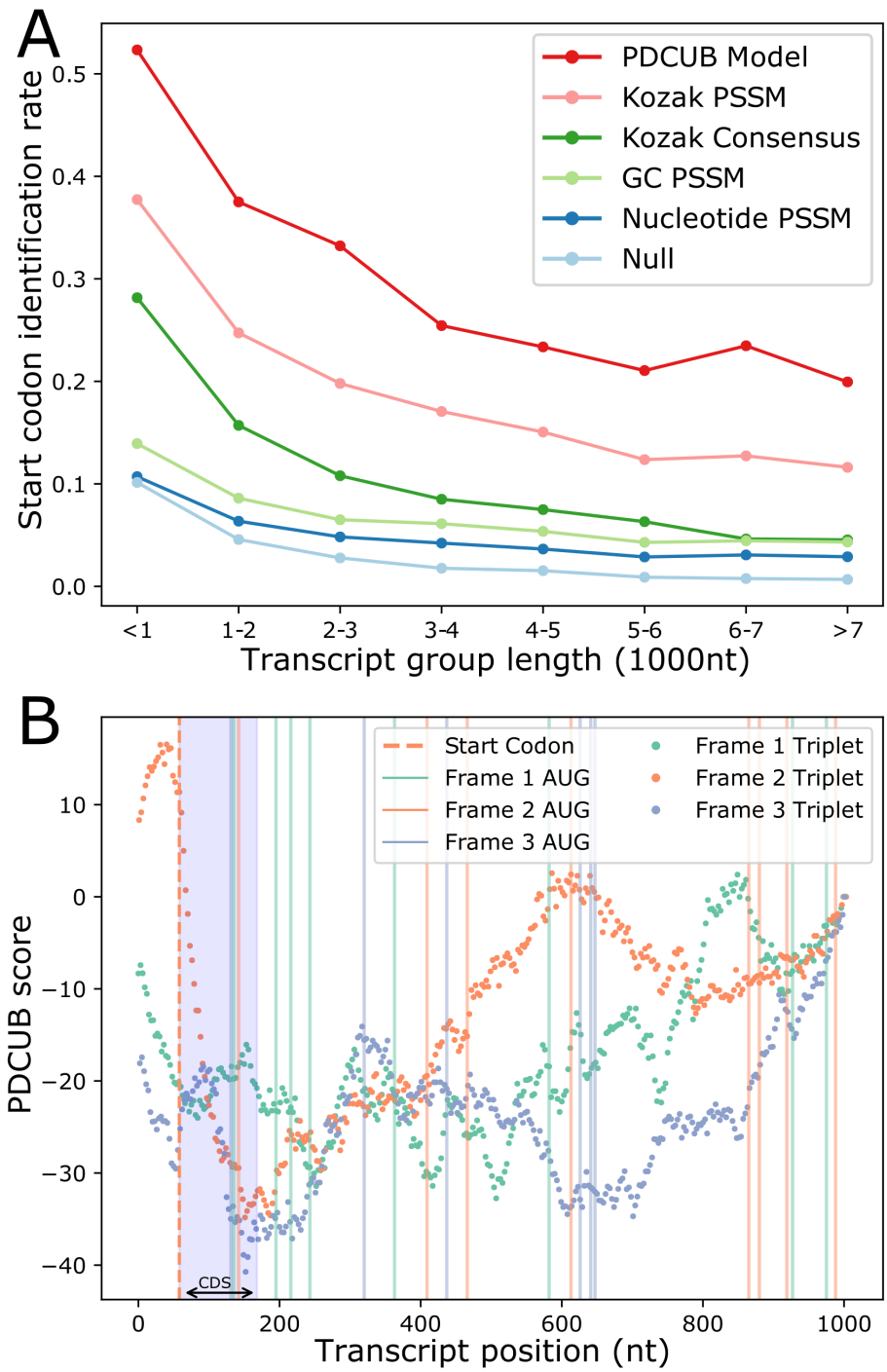

Figure 3. ***A.*** *The rate of start codon identification compared for different models. The curves show what fraction of start codons are identified per transcript, comparing the PDCUB score computed from a PSSM compared to the Kozak consensus sequence AUGGCC, a PSSM computed from the positions -10 to +2 relative to the start codon for all transcripts, a PSSM computed from the position-dependent GC content within CDS regions relative to global GC content, a PSSM computed from single-nucleotide frequencies in CDS regions relative to global nucleotide frequencies, a random selection model.* ***B.*** *The PDCUB score plotted as a function of position, starting at each codon, and plotted separately for each reading frame, for an example short (~1000nt) transcript. Blue shaded area beginning at the start codon indicates the CDS region.*

Functional enrichment in high-scoring transcripts

We identified a subset of transcripts with particularly high PDCUB scores relative to the distribution of scores observed over all AUG triplets (Supplementary Figure 4). To define a statistical threshold, we computed a p-value using the distribution of PDCUB scores over all AUG triplets, then computed a q-value using a Benjamini-Hochberg multiple test correction. We selected transcripts with a q-value less than 0.05, which we named our significant PDCUB set. We then examined whether transcripts in this set were enriched for biological functions or protein sequence patterns that might explain their high PDCUB scores.

Gene Ontology (GO) enrichment

We investigated whether the transcripts in our high- and low-PDCUB transcript sets were enriched or depleted for particular biological functions. A table of significant Gene Ontology (GO) terms is provided in Supplementary Table 1. All three GO hierarchies are examined, namely "cellular component", "molecular function", and "biological process". Enrichment of significant GO terms for high-PDCUB transcripts is visualized in a scatterplot (Figure 4A).

Growth and development. Among high-PDCUB transcripts, we observed a statistical enrichment of most developmental processes, with a corresponding depletion of these same processes among low-PDCUB transcripts. Examples include growth, development, differentiation, and morphogenesis of the nervous, circulatory, and musculoskeletal systems, as well as of various organs, glands, tissues, cells, and cellular components. Terms with the greatest enrichment were those relating to neurons, synapses, and axons, including over 1.5-fold enrichment of "regulation of synaptic plasticity" and "potassium ion transmembrane transport". The only exceptions to this trend were development-related terms involving skin, which were uniformly depleted among high-PDCUB transcripts but enriched among low-PDCUB transcripts.

Signaling and transport. Nearly all significant GO terms pertaining to signaling and transport are enriched among high-PDCUB transcripts, including terms related to synaptic signaling, cell-cell signaling, and signal transduction. The only exceptions to this trend occurred for signaling terms tied to immunological function, which are uniformly depleted among high-PDCUB transcripts. Meanwhile, signal and transport terms were mostly depleted among low-PDCUB transcripts.

Immune response and defense. All significant GO terms pertaining to immune response are depleted among high-PDCUB transcripts. This includes all terms relating to B cells and lymphocytes, which are immunological cells, as well as terms pertaining to phagocytosis and other modes of cellular and organismal defense.

Chemical stimulus. Significant GO terms pertaining to stimulus *detection* and *perception* are all depleted among high-PDCUB transcripts, including terms related to the sensory perception of smell. By contrast, significant GO terms related to stimulus *response* and *regulation* are all enriched among high-PDCUB transcripts. Among low-PDCUB transcripts, the exact opposite situation holds true, with uniform enrichment of GO terms involving stimulus detection and perception, and uniform depletion of terms involving stimulus response and regulation.

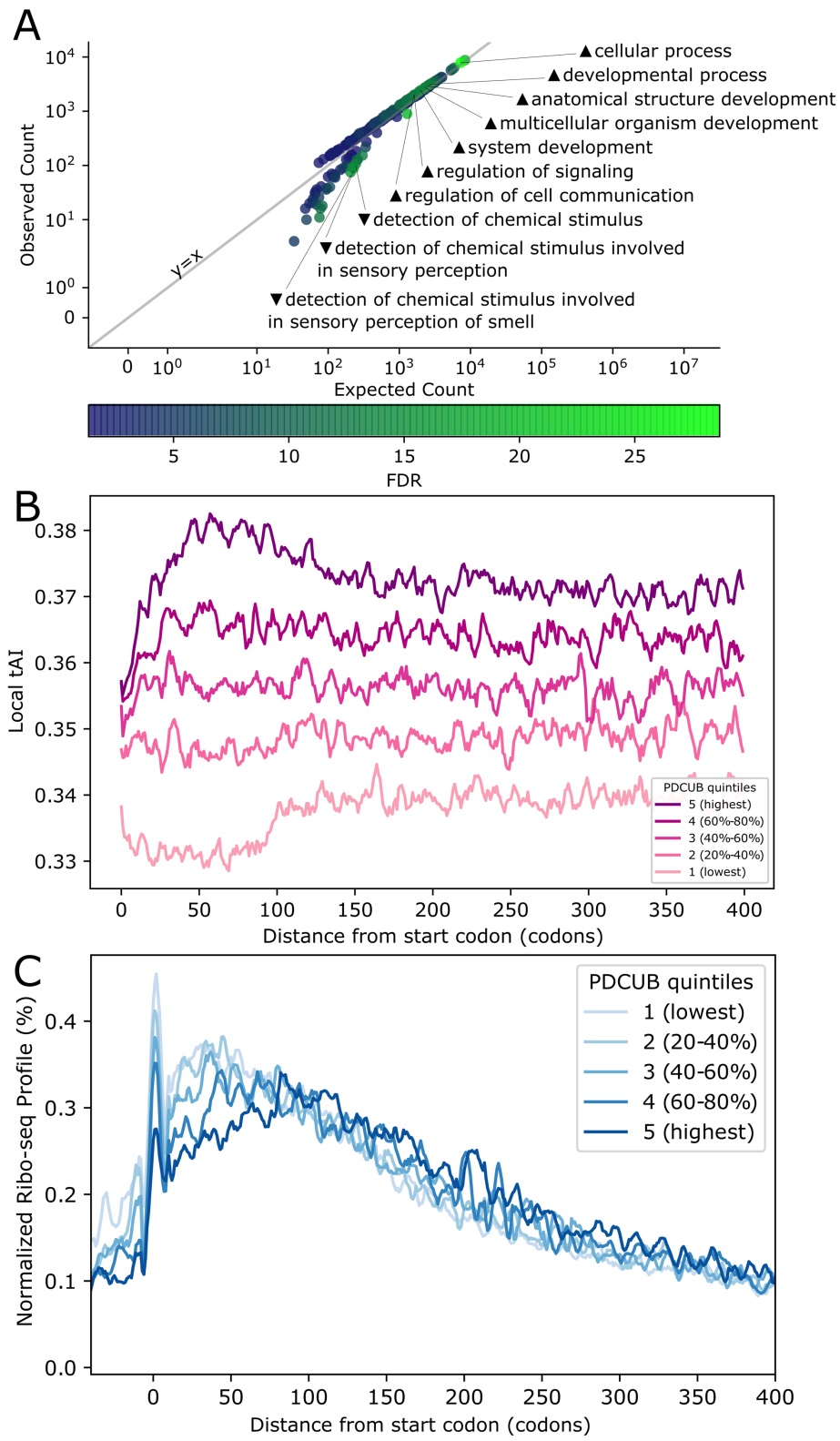

Figure 4: ***A.*** *Scatter plot representing statistically-enriched GO terms associated with transcripts that have statistically significant PDCUB scores compared to the scores of all AUGs within the human transcriptome. Each dot corresponds to an enriched GO term, the x-axis position corresponds to the expected number of occurrences, the y-axis represents the observe frequency within the PDCUB-enriched transcripts. Transcripts with the highest PDCUB scores are enriched. Triangles by the labels indicate enriched (pointed up) and depleted (down) GO terms.* ***B.*** *Local tAI profiles for transcripts in quintiles defined for PDCUB score.* ***C.*** *The shape of ribosomal occupancy, shown as the normalized Ribo-seq profile centered at the start codon at position 0. Read positions are A-shifted by 15-nt, and smoothed using a Savitzky-Golay filter.*

Signal Peptides

To investigate whether our results reflect the presence of nucleotide sequences coding for signal peptides, we used the signal-peptide prediction tools PrediSi [17] and SignalP [18] to predict which transcripts are likely to contain signal peptides. We performed a statistical test comparing the number of predicted signal peptides in the high-PDCUB transcripts to a background predicted rate for all protein-coding GENCODE transcripts not contained within the high-PDCUB set. We observe that 15-24% of proteins encoded by high-PDCUB transcripts are predicted to contain signal peptides, amounting to a 2-4% increase in predicted signal peptides among high-PDCUB transcripts relative to all other protein-coding transcripts in GENCODE (Supplementary Figure 5). We tested this difference via binomial test and found the increase to be statistically significant, with a p-value of $< 10^{-23}$. A Venn diagram of transcripts predicted by each approach to have a signal peptide, compared to transcripts with a high PDCUB score, is shown in Supplementary Figure 6.

Local tAI and CAI

We next sought to contextualize PDCUB against standard measures of translational efficiency, including the discovery of the translation initiation ramp [19]. We confirmed the presence of the translational ramp seen by Tuller *et al*, through a metagene plot of tAI against CDS position for the first 300 codons of the coding transcriptome (Supplementary Figure 7). We repeated this process with the transcriptome split into quintiles according to PDCUB scores, and found a clear, nonoverlapping spectrum of ascending average tAI values as we move from low-PDCUB transcripts to high-PDCUB transcripts (Figure 4B). This PDCUB-defined spectrum exhibits differently shaped local tAI curves, with the fifth quintile (highest PDCUB score) having the most pronounced translational ramp and lower quintiles having diminishing ramp size until the second quintile, which is flat with no ramp. More strikingly, the first quintile (lowest PDCUB score) exhibits a "translational valley", with tAI descending quickly from codons 0 to 30, before flattening out until codon 100, whereupon we observe a short ramp that is delayed relative to that of the highest PDCUB quintile. Finally, we plotted tAI against PDCUB score for the coding transcriptome and found poor correlation between the two parameters ($R^2 \approx 0.3$, Supplementary Figure 8).

We calculated local tAI values for CDS regions corresponding to the first and second windows of length w codons, for $w = 30$ and $w = 50$, for all transcripts. The midpoints of these two sets of paired windows correspond, respectively, to the lower and upper bounds reported for the size of the translation initiation ramp and should thus demark a boundary between lower and higher translational efficiency signals, which should be apparent when plotting the local tAIs for the second window against the first. No such boundary was apparent in our scatter plot, regardless of the PDCUB of the corresponding transcripts (Supplementary Figure 9). However, we did see a clear gradient of PDCUB scores across the transcript distribution in both codon windows, with low-PDCUB transcripts concentrated at low tAI values, and high-PDCUB transcripts concentrated at high tAI values.

As an alternative approach to understanding how PDCUB compares to translational efficiency, we performed a similar set of analyses with CAI scores computed from window-specific codon frequencies (local CAI). This revealed a distinct subset of low-PDCUB transcripts in which the local CAI of the first 30-50 codons in the CDS (corresponding to the translation initiation ramp) is substantially lower than that of the immediately subsequent 30-50 codons (corresponding to the translation steady state) (Supplementary Figure 10).

Ribo-seq profiles

Given the observation that higher ribosomal density is associated with lower translational speed [20], we examined whether transcripts with high PDCUB scores show distinct trends in ribosomal occupancy compared to transcripts with low PDCUB scores. Figure 4C shows the normalized ribosomal occupancy relative to the start codon for each PDCUB quintile. We observed a spectrum of ribosome profiles where the first quintile (lowest PDCUB) has the highest relative occupancy at initiation and the fifth quintile (highest PDCUB) has the lowest relative occupancy.

DISCUSSION

We observe position-dependent trends when visualizing codon usage across all protein-coding regions in the human transcriptome. Additionally, over one third of ORFs in individual protein-coding transcripts exhibit a statistically significant PDCUB when compared to reading frames that follow non-start AUGs. We have shown that start codons can be distinguished from other AUGs across reading frames solely by using a score that quantifies PDCUB for the 300 nt immediately following the AUG, thus further validating that PDCUB is a distinguishing feature of many start codons. We also observe a substantial difference between the ribosomal occupancy for high-PDCUB transcripts compared to those with low PDCUB. These observations are consistent with the hypothesis that human translational machinery, including ribosomes and translation initiation factors, rely broadly upon positionally dependent sequence patterns to direct and regulate multiple aspects of translation [21].

As discussed in previous studies in yeast [13], we tested whether our observed PDCUB pattern could be explained by signal peptides, and found a 2-4% enrichment of signal peptides in proteins encoded by high-PDCUB transcripts. This low enrichment suggests that while sequence features that code for signal peptides may contribute to PDCUB score, they are not a driving mechanism of PDCUB. It is also worth noting that while signal peptides are defined by an amino acid sequence, PDCUB describes position-dependent patterns at the nucleotide level in addition to the triplet level, and thus may capture translationally important patterns that position-dependent amino acid signals alone cannot completely resolve.

Previous research has shown that GC-rich codons have greater translational efficiency [22]. We observe that the pattern of PDCUB across transcripts shows enrichment of GC content in codons at the beginning of coding regions relative to downstream segments. Simultaneously, we observe lower ribosomal occupancy over the region in which PDCUB is calculated (1-300 nt downstream of the start codon). Therefore, we speculate that observed PDCUB patterns may lead to greater efficiency of translation initiation and possibly also translation elongation.

Selection for GC-rich, translationally efficient coding regions may encourage the development of PDCUB patterns, or vice versa. However, the fact that PDCUB outperforms GC content in the identification of start codons suggests that GC content is not the whole story, and higher-order sequence patterns may be at play.

We observed that PDCUB defines a spectrum of translational efficiency, as observed by both ribosome occupancy and local tAI curves. Previous studies have been conflicted about the presence of a translational ramp in mammals, with some reporting no ramp [23] and others reporting a distinct ramp [19]. We find that both results are true depending on where one looks within the PDCUB spectrum, with high-PDCUB transcripts having a strong initial ramp, mid-PDCUB transcripts having a weaker ramp or no ramp, and low-PDCUB transcripts exhibiting an initial translational valley. In particular, high-PDCUB transcripts seem to exhibit the same translational ramp described elsewhere [19, 24], and this translational ramp corresponds to reduced ribosomal occupancy in the early regions of those same transcripts.

Both high- and low-PDCUB transcripts are biased toward reduced local tAI *and* reduced local CAI in the first 30-50 codons, relative to the immediately subsequent 30-50 codons. This makes sense intuitively, as both tAI and CAI are intended as measures of translational efficiency and should be lower during a ramp than on the plateau following the ramp, or during a valley than during the plateau following the valley. Interestingly, low-PDCUB transcripts show a larger gap between first and second codon windows than do high-PDCUB transcripts. As PDCUB score differences are greater between codon windows in the highest quintile (which has a strong ramp) than the lowest quintile (which has a valley), this is further evidence that PDCUB captures important information regarding translational efficiency that is not captured by CAI or tAI alone. These results serve as a basis for understanding the complex relationship between PDCUB, translational ramps, and other metrics of codon bias, including the extent to which they might be complementary or competing factors in translational efficiency.

Importantly, PDCUB defines transcript subpopulations that correspond to transcript function. In our analysis of GO terms, we observe that functions involved in growth and development, including for nervous, cardiovascular, and musculoskeletal systems, are highly enriched among high-PDCUB transcripts, as are growth terms at the organ, tissue, cellular, and subcellular levels. Among low-PDCUB transcripts, these same terms are depleted. This may be explained by the fact that regulation of organismal development and cellular differentiation require precise timing [25, 26], such that these processes are forced to rely on PDCUB to enhance translational efficiency by improving the timing of regulatory cascades. This may be true particularly at the initial translation initiation step, thereby accounting for the pronounced translational ramp found in high-PDCUB transcripts. We further note that these developmental pathways are part of precise regulatory functions that require yet more exquisite tuning of translational control [27, 28].

The only exceptions to the trends described above are developmental terms involving skin, which are universally depleted among high-PDCUB transcripts but enriched among low-PDCUB transcripts. This set of exceptions may be tied to the prominent role of skin in human immune function relative to other organs and tissues, as skin comprises the body's first and primary barrier against pathogens. More generally, we observe that transcripts related to immune response, as well as those governing detection and perception of chemical stimuli, are statistically depleted among high-PDCUB transcripts and enriched among low-PDCUB transcripts.

Immune response and chemical detection pathways have in common high genetic copy number and sequence variation [29]. Moreover, there is an abundance of pseudogenes observed for human odorant receptors and other genes responsible for detection of chemical stimuli [30], as well as immune genes [31, 32]. These observations may suggest a lower selective pressure on individual protein-coding genes due to the robustness of these pathways on the whole, conferred by high copy number and a need to constantly adapt to changing chemical [33] and pathogen [34-36] environments. We posit that genes within these pathways might sample a broader range of codon ordering compared to genes involved in the development of complex vital organs. Meanwhile, genes tied to chemical stimulus *response* and *regulation* show a reversed trend relative to detection and perception, with enrichment among high-PDCUB transcripts and depletion among low-PDCUB transcripts. This trend reversal may stem from a need for precise timing among regulatory and response genes, as described for developmental genes above, while detection and perception genes rely on high copy number and low translational control to account for the complex suite of chemicals that must be adapted to in the human sensory environment.

MATERIALS AND METHODS

The dataset used was the GENCODE Release 34 protein-coding transcript sequences FASTA [15]. We filtered out all transcripts in which the length of the labelled coding region was not a multiple of three, lacked an in-frame start codon, or lacked an in-frame stop codon. This dataset was used for all of our analyses.

Codon frequencies relative to transcript sequence position

Codons following the start codon were counted in bins of size 60 (20 nucleotides) for a total of 50 bins. To keep the bin size constant, the probability of finding a given codon in a bin was calculated, and subsequently used to derive a mean and standard deviation for each codon. The z-score of a given sense codon in a given bin in Figure 1A was calculated by

$$z_{c,b} = \frac{p_{c,b} - \mu_c}{\sigma_c}$$

where $p_{c,b}$ is the probability of a codon in a given bin, $\mu_c$is the mean of the codon across all bins, and $\sigma_c$is the standard deviation of the codon across all bins.

Hierarchical clustering of codons based on z-score profiles

To obtain the dendrogram of z-score profiles, a distance matrix was first calculated wherein each row and column corresponds to a single codon. A cell at row *m* and column *n* would thus correspond to the distance between the z-score profiles of codons *m* and *n.* The distance between two codons was calculated based on the sum of squares of the differences between z-scores:

$$D_{m,n} = \sum_{b}^{50} (z_{m,b} - z_{n,b})^2$$

where $z_{m,b}$ and $z_{n,b}$ are respectively the z-scores of codons *m* and *n* at bin b.

The dendrogram was computed using python's scipy.cluster.hierarchy.linkage method [37], which performs clustering given a distance matrix such as the one calculated above. The method used for calculating distance between two clusters *u* and *v* was the 'average' method

$$d(u, v) = \sum_{i,j} \frac{d(u[i], v[j])}{|u| * |v|},$$

which describes the distance between two clusters of codon profiles over positions i and j relative to the cardinalities of clusters u and v.

Scoring models for predicting start codons

Several scoring models were developed to predict the location of start codons, in order to determine the effect of position dependence in the human transcriptome. These models were based on a weight matrix, or position specific-scoring matrix (PSSM), where the rows correspond to codons or nucleotides and the columns correspond to positions after the start codon. In general, the PSSM values were calculated as log likelihoods based on a background model.

<u>PDCUB PSSM</u>

The weights in the PDCUB PSSM were calculating according to the equation

$$W_{cb} = log_2 \left(\frac{f_{cb}}{p_c} \right)$$

where f_{cb} is the frequency of codon c at bin b, and p_c refers to the background probability of codon c. Bins b of size w (in codons) with start positions s_b comprise a series $\{[s_1, s_1 + w], [s_2, s_2 + w], \dots [s_N, s_N + w]\}$ up to N bins for each putative CDS. We define a putative CDS region as series of triplets after the AUG, $C = (c_1, c_2, \dots, c_n)$. The PDCUB score $S(C)$ is computed for a given CDS according to

$$\mathrm{S(C)} = \sum_{b=1}^{N} \sum_{i=s_b}^{w} W_{c_i,b}$$

with PDCUB models computed at bin sizes of w = 1, 2, 3, 4, 5, 10, 20, 25, 50, and 100 codons. After examining the percentage of transcripts in which the true start codon was the highest scoring AUG, it was determined that a bin size of 3 nucleotides was optimal.

<u>Kozak consensus sequence</u>

The Kozak consensus sequence is defined as (gcc)gccRccAUGG, where R corresponds to A or G, and position +1 corresponds to the A in the AUG. Positions -3 (the R) and +3 (the G after the AUG) are particularly important, and upper-case indicates highly conserved bases [16]. We gave positions +3 and -3 a weight of 6, and the remaining positions a weight of 1. The score of +6 for each highly conserved position ensures that correctly matching one highly conserved position outweighs any number of less-conserved matches. Transcripts were scored based on the number of mismatches the surrounding nucleotides had with the consensus sequence, with a maximum score being 17 (no mismatches). The leading (gcc) was subtracted from this score, so that codons containing the matching sequences were rewarded, but codons without them were not penalized.

Kozak PSSM

The Kozak PSSM model was implemented in a manner similar to that for PDCUB, with a few key differences. A PSSM was built by computing the frequencies $f_{b,i}$ for each nucleotide b at position i relative to the start codon across all protein-coding transcripts. The PSSM is then computed using this frequency relative to the global single-nucleotide frequencies p_b, as shown below.

$$W_{b,i} = log_2\left(\frac{f_{b,i}}{p_b}\right)$$

Therefore, the Kozak PSSM differs from the PDCUB PSSM because the former was based on nucleotide counts instead of codon counts. The positions in the Kozak PSSM correspond to the ten nucleotides before the start codon, which is the location of the Kozak consensus sequence. All positions in the transcript were thus scored by taking the sum of the weights of ten preceding nucleotides.

GC content PSSM

The GC content PSSM has two rows, one corresponding to 'S' nucleotides {G,C} and the other to 'W' nucleotides {A,U}. Nucleotides were counted into either of those categories in a manner similar to the nucleotide PSSM used for the Kozak weight matrix.

Statistical Null Model

The null model was calculated to determine the apparent extent of random chance involved in an AUG being the true start codon in a given transcript. All AUG triplets in a transcript were scored, and an AUG was randomly selected as the start codon. If the selected AUG had the highest score, then the model was considered successful for that transcript.

Theoretical Model

If the length of a given transcript is L, then there are $L - 2$ possible positions in which a triplet can exist when allowing all three possible reading frames. Moreover, the codon at that given position must be an AUG, of which there is a 1/64 chance. Therefore, the total number of expected AUGs in a transcript is

$$\langle N_{AUG} \rangle = (L-2)(\frac{1}{64})$$

The theoretical probability that an AUG selected a random is the true start codon can thus be modeled by

$$\frac{1}{\langle N_{AUG} \rangle} = \frac{64}{L-2}$$

Cross-validation methods

The dataset of all valid sequences was split into three separate sets. In all cases, 80% of the data (about 55,000 transcripts) were used to obtain PSSMs of the model in question. In most cases, the remaining 20% (about 14,000 transcripts) were scored with the calculated PSSM. In the case of PDCUB, where multiple bin sizes were tested, 10% of the data was used to determine the bin size that produced the highest-scoring AUG triplets, and 10% was left for scoring.

Information content of bins within a transcript

The information content IC of position i in a transcript was calculated as follows:

$$IC(i) = H_f - H_p(i)$$

where H_f is the Shannon entropy of the global triplet frequencies (background model) of the transcript, and $H_p(i)$ is the Shannon entropy of the position-dependent codon frequencies at position i. The background entropy is calculated as

$$H_f = -\sum_c f_c log_2(f_c)$$

where f_c is the global frequency of triplet *c.* The product of the global frequency and the log of the global frequency are computed for all triplets and then added together. The Shannon entropy of the position-dependent codon frequencies is calculated as

$$H_p(i) = -\sum_c p_{c,i}\, log_2(p_{c,i})$$

where $p_{c,i}$ is the probability of codon c at position i.

Determination of significant high-scoring AUG triplets

After generating a list of transcripts and their highest-scoring AUG triplet based on the PDCUB model, the p-values for each AUG score were then calculated. These values were filtered with the Benjamin-Hochberg correction to produce a list of statistically significant genes ($\alpha < 0.05$). All scores were sorted by p-value and accordingly ranked (the lowest p-value would have a rank of 1, the next lowest a rank of 2, and so on). A corresponding q-value was then calculated with

$$q_r = \frac{P_r * N}{r}$$

where *r* is the rank and *N* is the total number of scores. All transcripts with a rank that was lower than the highest rank r with q_r below the significance level were deemed statistically significant.

Analyzing the contribution of signal peptides

Prediction of signal peptides in human proteome

The mRNA content of the human transcriptome was retrieved from GENCODE, converted to protein sequence format, and fed to the online tools PrediSi [17] and SignalP-5.0 [18] for prediction of signal peptides. Both tools were accessed through their public-facing online interfaces using default program parameters. In both cases, transcripts from the significant PDCUB set were separated from the remainder of the GENCODE set.

Statistical enrichment of signal peptides in significant PDCUB set

Densities and counts for significant PDCUB and complement sets were compared in two batches: one for the PrediSi scores, and another for the SignalP scores. In both cases, the significant PDCUB set was tested for significant enrichment of signal peptides relative to the complement set, which we defined to contain the background signal peptide rate. Significance was determined via binomial test.

Overlap between significant PDCUB and signal peptides

Transcripts from the significant PDCUB set were checked for overlap with all GENCODE protein-coding transcripts predicted to have a signal peptide. This was done for two sets of transcripts: GENCODE transcripts that scored positively for signal peptides using PrediSi, and those that scored positively for signal peptides using SignalP-5.0. All three resulting sets (significant PDCUB, PrediSi GENCODE, and SignalP-5.0 GENCODE) were visualized together as a Venn diagram using matplotlib.

Ribosome profiling

Ribosome profiling of HEK293 cells was downloaded from GSE accession number GSE102720 [38]. Ribo-seq reads were aligned to protein-coding transcripts using Bowtie [39] with the parameters "-v 2 -a -S --strata --best --norc -m 200" after trimming the adaptors with Cutadapt [40]. Read 5′ ends were shifted relative to the start of the CDS and an additional 15 nucleotides to align to the A site [41]. Normalization of each individual transcript was performed by dividing the sum of the reads at each nucleotide position by the total number of reads mapping up to 500

codons from the start codon for that transcript, and converted to a percentage. The sum at each position of these normalized transcripts was divided by the total number of transcripts to create an average profile. The average curves for each quintile were smoothed using a Savitzky-Golay filter [42] with a window size of 19 and a polynomial degree of 4 using the scipy implementation [37].

Gene Ontology Enrichment

ENSEMBL IDs for high-PDCUB and low-PDCUB transcripts were analyzed using the GO Enrichment Analysis tool, which is part of the web-accessible Gene Ontology Resource (http://geneontology.org/) [43-45]. Tables of statistically significant GO terms were retrieved for both sets of transcripts. The tables are provided as excel files in Supplementary Table 1. Significant GO terms were those with unexpected enrichment or depletion among high- or low-PDCUB transcripts, when a Benjamini-Hochberg multiple test correction with an FDR of 0.05 was applied.

Calculating local CAI and tAI

Sets of local relative adaptiveness (RA) and CAI values were calculated for all protein-coding GENCODE human transcripts. Whereas RA weights are traditionally defined by one unique set of values for a given transcriptome, and are derived from the total codon composition across the full lengths of every CDS, our local RA weights were computed for a particular region r of every CDS. These local RA values were used as input weights w_{rc} to calculate local CAI values for the same region r and codon c of a single transcript. For example, when examining the local CAI of the first 50 codons in a given CDS, our local RA was derived exclusively from the first 50 codons of every CDS. The local RA was calculated as

$$w_{rc} = \frac{f_{rc}}{max(f_{rc'})}$$

where f_{rc} is the frequency of codon c in region r and $f_{rc'}$ is the frequency of codon c' in region r, where c and c' are both synonymous codons for the same amino acid. From this, the local CAI was calculated as

$$CAI_r = \left(\prod_{k=1}^{l_r} w_{rc_{kr}}\right)^{1/l_r}$$

where l_r is the length of the transcript region, and c_{kr} is the codon defined by the kth triplet in that region.

To calculate local tAI, we retrieved human codon tAI weights from Table S1 of Tuller *et al.* [19] and computed a similar geometric mean for the same regions as above. Quintile tAI figures were generated by splitting GENCODE transcripts into five parts and plotting average local tAI for each quintile as a function of codon position, using a sliding window five codons in length.

Code for this project is available at https://github.com/hendrixlab/PDCUB.

References

1. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S: **The landscape of long noncoding RNAs in the human transcriptome**. *Nature genetics* 2015, **47**(3):199-208.
2. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES: **Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins**. *Cell* 2013, **154**(1):240-251.
3. Bánfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, Kundaje A, Gunawardena HP, Yu Y, Xie L: **Long noncoding RNAs are rarely translated in two human cell lines**. *Genome research* 2012, **22**(9):1646-1657.
4. Ji Z, Song R, Regev A, Struhl K: **Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins**. *elife* 2015, **4**:e08890.
5. Li L-J, Leng R-X, Fan Y-G, Pan H-F, Ye D-Q: **Translation of noncoding RNAs: focus on lncRNAs, pri-miRNAs, and circRNAs**. *Experimental cell research* 2017, **361**(1):1-8.
6. Makarewich CA, Olson EN: **Mining for micropeptides**. *Trends in cell biology* 2017, **27**(9):685-696.
7. Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, McAnally JR, Kasaragod P, Shelton JM, Liou J, Bassel-Duby R: **A micropeptide encoded by a putative long noncoding RNA regulates muscle performance**. *Cell* 2015, **160**(4):595-606.
8. Zhang H, Zhang L, Li Z, Liu K, Liu B, Mathews DH, Huang L: **LinearDesign: Efficient Algorithms for Optimized mRNA Sequence Design**. *arXiv preprint arXiv:200410177* 2020.
9. Kozak M: **Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs**. *Nucleic acids research* 1984, **12**(2):857-872.
10. Wright G, Rodriguez A, Li J, Clark PL, Milenković T, Emrich SJ: **Analysis of computational codon usage models and their association with translationally slow codons**. *PloS one* 2020, **15**(4):e0232003.
11. Sharp PM, Li W-H: **The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications**. *Nucleic acids research* 1987, **15**(3):1281-1295.
12. Hockenberry AJ, Sirer MI, Amaral LAN, Jewett MC: **Quantifying position-dependent codon usage bias**. *Molecular biology and evolution* 2014, **31**(7):1880-1893.
13. Villada JC, Brustolini OJB, Batista da Silveira W: **Integrated analysis of individual codon contribution to protein biosynthesis reveals a new approach to improving the basis of rational gene design**. *DNA Research* 2017, **24**(4):419-434.
14. Hill ST, Kuintzle R, Teegarden A, Merrill III E, Danaee P, Hendrix DA: **A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential**. *Nucleic acids research* 2018, **46**(16):8105-8113.
15. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J: **GENCODE reference annotation for the human and mouse genomes**. *Nucleic acids research* 2019, **47**(D1):D766-D773.
16. Kozak M: **An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs**. *Nucleic acids research* 1987, **15**(20):8125-8148.
17. Hiller K, Grote A, Scheer M, Münch R, Jahn D: **PrediSi: prediction of signal peptides and their cleavage positions**. *Nucleic acids research* 2004, **32**(suppl_2):W375-W379.

18. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H: **SignalP 5.0 improves signal peptide predictions using deep neural networks**. *Nature biotechnology* 2019, **37**(4):420-423.
19. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y: **An evolutionarily conserved mechanism for controlling the efficiency of protein translation**. *Cell* 2010, **141**(2):344-354.
20. Ingolia NT: **Ribosome profiling: new views of translation, from single codons to genome scale**. *Nature reviews genetics* 2014, **15**(3):205-213.
21. Mitarai N, Pedersen S: **Control of ribosome traffic by position-dependent choice of synonymous codons**. *Physical biology* 2013, **10**(5):056011.
22. Hia F, Yang SF, Shichino Y, Yoshinaga M, Murakawa Y, Vandenbon A, Fukao A, Fujiwara T, Landthaler M, Natsume T: **Codon bias confers stability to human mRNA s**. *EMBO reports* 2019, **20**(11):e48220.
23. Ingolia NT, Lareau LF, Weissman JS: **Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes**. *Cell* 2011, **147**(4):789-802.
24. Verma M, Choi J, Cottrell KA, Lavagnino Z, Thomas EN, Pavlovic-Djuranovic S, Szczesny P, Piston DW, Zaher HS, Puglisi JD: **A short translational ramp determines the efficiency of protein synthesis**. *Nature communications* 2019, **10**(1):1-15.
25. Bolouri H, Davidson EH: **Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics**. *Proceedings of the National Academy of Sciences* 2003, **100**(16):9371-9376.
26. Daou R, Beißbarth T, Wingender E, Gültas M, Haubrock M: **Constructing temporal regulatory cascades in the context of development and cell differentiation**. *PloS one* 2020, **15**(4):e0231326.
27. Fujii K, Shi Z, Zhulyn O, Denans N, Barna M: **Pervasive translational regulation of the cell signalling circuitry underlies mammalian development**. *Nature communications* 2017, **8**(1):1-13.
28. Hershey JW, Sonenberg N, Mathews MB: **Principles of translational control: an overview**. *Cold Spring Harbor perspectives in biology* 2012, **4**(12):a011528.
29. De Smith A, Walters R, Froguel P, Blakemore A: **Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease**. *Cytogenetic and genome research* 2008, **123**(1-4):17-26.
30. Spehr M, Munger SD: **Olfactory receptors: G protein-coupled receptors and beyond**. *Journal of neurochemistry* 2009, **109**(6):1570-1583.
31. Vargas-Madrazo E, Almagro JC, Lara-Ochoa F: **Structural repertoire in VHpseudogenes of immunoglobulins: comparison with human germline genes and human amino acid sequences**. *Journal of molecular biology* 1995, **246**(1):74-81.
32. Ortutay C, Siermala M, Vihinen M: **Molecular characterization of the immune system: emergence of proteins, processes, and domains**. *Immunogenetics* 2007, **59**(5):333-348.
33. Niimura Y: **Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents**. *Human genomics* 2009, **4**(2):1-12.
34. Broecker F, Moelling K: **Evolution of immune systems from viruses and transposable elements**. *Frontiers in microbiology* 2019, **10**:51.

35. Cooper MD, Alder MN: **The evolution of adaptive immune systems**. *Cell* 2006, **124**(4):815-822.
36. Quintana-Murci L: **Human immunology through the lens of evolutionary genetics**. *Cell* 2019, **177**(1):184-199.
37. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J: **SciPy 1.0: fundamental algorithms for scientific computing in Python**. *Nature methods* 2020, **17**(3):261-272.
38. Iwasaki S, Iwasaki W, Takahashi M, Sakamoto A, Watanabe C, Shichino Y, Floor SN, Fujiwara K, Mito M, Dodo K: **The translation inhibitor rocaglamide targets a bimolecular cavity between eIF4A and polypurine RNA**. *Molecular cell* 2019, **73**(4):738-748. e739.
39. Langmead B: **Aligning short sequencing reads with Bowtie**. *Current protocols in bioinformatics* 2010, **32**(1):11.17. 11-11.17. 14.
40. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *EMBnet journal* 2011, **17**(1):10-12.
41. Guo H, Ingolia NT, Weissman JS, Bartel DP: **Mammalian microRNAs predominantly act to decrease target mRNA levels**. *Nature* 2010, **466**(7308):835-840.
42. Savitzky A, Golay MJ: **Smoothing and differentiation of data by simplified least squares procedures**. *Analytical chemistry* 1964, **36**(8):1627-1639.
43. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: **Gene ontology: tool for the unification of biology**. *Nature genetics* 2000, **25**(1):25-29.
44. **The Gene Ontology resource: enriching a GOld mine**. *Nucleic Acids Research* 2021, **49**(D1):D325-D334.
45. Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD: **PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API**. *Nucleic Acids Research* 2021, **49**(D1):D394-D403.